View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Edinburgh Research Explorer

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

# MINIMIZING MAKESPAN IN A MULTICLASS FLUID NETWORK WITH PARAMETER UNCERTAINTY

**Citation for published version:**
Buke, B, Hasenbein, J & Morton, D 2009, 'MINIMIZING MAKESPAN IN A MULTICLASS FLUID NETWORK WITH PARAMETER UNCERTAINTY' Probability in the Engineering and Informational Sciences, vol. 23, no. 3, pp. 457-480. DOI: 10.1017/S026996480900028X

**Digital Object Identifier (DOI):**
10.1017/S026996480900028X

**Link:**
Link to publication record in Edinburgh Research Explorer

**Document Version:**
Publisher's PDF, also known as Version of record

**Published In:**
Probability in the Engineering and Informational Sciences

**General rights**
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 05. Apr. 2019

*Probability in the Engineering and Informational Sciences*, **23**, 2009, 457–480. Printed in the U.S.A.
doi:10.1017/S026996480900028X

# MINIMIZING MAKESPAN IN A MULTICLASS FLUID NETWORK WITH PARAMETER UNCERTAINTY

Burak Büke, John J. Hasenbein, and David P. Morton
*Graduate Program in Operations Research & Industrial Engineering*
*Department of Mechanical Engineering*
*The University of Texas at Austin*
*Austin, TX 78712-0292*
*E-mail: {bukeb, jhas, morton}@mail.utexas.edu*

We introduce and investigate a new type of decision problem related to multiclass fluid networks. Optimization problems arising from fluid networks with known parameters have been studied extensively in the queueing, scheduling, and optimization literature. In this article, we explore the makespan problem in fluid networks, with the assumption that the parameters are known only through a probability distribution. Thus, the decision maker does not have complete knowledge of the parameters in advance. This problem can be formulated as a stochastic nonlinear program. We provide necessary and sufficient feasibility conditions for this class of problems. We also derive a number of other structural results that can be used in developing effective computational procedures for solving stochastic fluid makespan problems.

## 1. INTRODUCTION

In this article, we analyze the structural properties of a *stochastic fluid model.* Our model is an extension of the classical multiclass fluid model that has been studied in a number of articles over the last 20 years (see [10,11] for background on such models). In a multiclass fluid model, fluids of various types enter a network at given, constant rates. Fluid is then processed at a station at a given rate and then routed to another station for processing or it might leave the network. A standard optimization problem in such networks is to drain the network of fluid in the least amount of time, given an initial fluid inventory. This is sometimes called the *fluid makespan problem* or the *clearing time problem.* A related problem, known as the *fluid holding cost problem*,

© 2009 Cambridge University Press

is to drain the network with the lowest cost, where the cost is some function of the fluid levels in the network. The former problem is a relatively simple optimization problem, and computing its solution only requires inversion of the fluid routing matrix. The latter problem is much more difficult in general and falls into the class of separated continuous linear programs (see, e.g., [1]). These problems have received attention in a number of studies, a short list of which includes [4,7,10,28,29,34,35].

Our model is an extension of the fluid model described earlier in that we allow some of the parameters, specifically the fluid arrival and processing rates, and the initial inventory to be random vectors. Such a network is referred to as a stochastic fluid model. We formally describe the model in Section 3, but we also give an informal description here in order to contrast our model with others that have been analyzed in the literature. Our problem is a stochastic optimization problem with a relatively simple decision structure. Before time 0, the decision maker must choose a set of "allocation percentages" $v_k$, which determine what percentage of a server's capacity will be devoted to class $k$ fluid (assuming there is a sufficient amount of fluid to be worked on). These percentages are then fixed once and for all. At time 0, when the system begins operation, a realization of the stochastic parameters is revealed, and the system then operates under that realization and the allocation percentages $v_k$ that were chosen. The controller's goal is to choose the allocations $v_k$ in a manner that will minimize the expected draining time of the system. Hence, our problem is a stochastic version of the fluid makespan problem. We focus on the makespan objective due to its relative simplicity, although some of our results also apply to more general objective functions.

We view this model as a useful approximation of reality in systems where at least the following characteristics are present: (1) the dynamical aspects of the system are well approximated by a multiclass fluid model, in particular the possible discrete nature of the system and small time scale stochastic fluctuations are well represented by a deterministic, continuous model; (2) the stochastic behavior of some structural parameters of the system are dominant in terms of system behavior; and, (3) the decision maker is constrained in the sense that some irreversible training or allocation decisions must be made before the stochastic structural parameters can be measured. Systems in a number of different application areas do have these characteristics and we mention just a few. In a recent article, Harrison and Zeevi [20] presented a compelling argument for using a model of a similar nature in call center applications. In particular, in their model, incoming calls are approximated on a local time scale by a deterministic fluid process. However, over longer time scales, they assume that the incoming call rates have some stochastic variability that is the dominant random factor. Finally, they posit that call center staffing decisions must be made *before* the incoming call rates are known. Thus, their modeling framework for call centers coincides with our modeling regime.

In semiconductor wafer manufacturing, the dynamics of the manufacturing process can often be well approximated by a multiclass fluid model when there is a high production volume in the wafer fab. The dominant uncertainties in a wafer fab are usually in terms of demand rates (i.e., lot arrival rates) and the availability of critical equipment, due to unscheduled downtimes. In some cases, machine purchases, reticle availability, and setups constrain the local time scale decisions of machines allocated to different products. Hence, this application domain provides another motivation for our model.

A number of different stochastic fluid models have been introduced in the literature. There is a large body of work on models related to the classical Anick–Mitra–Sondhi [2] stochastic fluid model. In those models, generally speaking, service rates are deterministic and arrival rates vary according to an underlying Markov chain (i.e., the arrival process is *Markov modulated*). The controller's job is usually to determine which fluid classes to serve at any given time and how much of each fluid type to admit to the system in order to minimize a cost function. For modeling of manufacturing systems, both the incoming fluid rate and the processing rates at a server might be allowed to vary according to some stochastic process. Again, in those models the controller might be allowed to control both admissions to the system and the servers' time allocations. For different approaches to these problems, see, for example, [5,6,19,32]. Overviews of stochastic fluid models used in the manufacturing and telecommunications application appear in [25,30]. A recent book by Meyn [27] provides a comprehensive overview of the relation between the control of fluid models and scheduling complex stochastic networks.

The models mentioned earlier differ from ours in terms of the decision structure in an important way. In the models of previous work, one usually has the freedom to modify allocation or admission decisions as soon as a change in the system parameters is observed, and in this sense, the decision structure is that of real-time control. We do not have the luxury of being able to quickly adapt to changes in the system's parameters. Rather, the controller must commit to a decision ahead of time and then live with the consequences of that decision regardless of the realization of the stochastic parameters (i.e., the decision structure of our model is that of a time-static stochastic program). This structure is reasonable when the controller must make a decision concerning the design of the system. For example, the number of dedicated servers (e.g., the number of trained personnel) to accomplish certain tasks should be decided before the system starts running and it might be too costly, or logistically impossible, to change this decision after the system parameters are observed. Another example, where it is not possible to modify the decision after realizing the parameters, is signal control for heavy traffic in urban areas. The controller must decide ahead of time on the duration of red and green lights at each intersection without observing the actual flow in the system. In this case, induction sensors at intersections do not provide sufficient real-time information on the flow at every intersection.

Perhaps the model closest to ours in spirit is in the aforementioned article [20]. Using the motivation for parameter uncertainty in [20], we study the makespan problem for fluid networks that has been studied in Weiss [34] and Dai and Weiss [12]. Harrison and Zeevi [20] studied a more general server structure than ours, as their network has flexible servers; that is, a fluid class can be served by more than one station in the network. However, their network structure is simpler, since they only consider "one pass" networks in which fluid visits only one server and then departs the network. Also in contrast to [20], we focus on the structural aspects of the stochastic programming problem that arises. As in [20], Atlason, Epelman, and Henderson [3] optimized staffing levels at a call center but they also addressed the combinatorial problem of constructing employee shifts while using a simulation model to estimate the center's performance. Gürkan [18] selected constrained buffer sizes to optimize throughput in a fluid tandem queueing network with random machine failures, again using simulation to estimate steady-state throughput. The recent articles of Iyengar and Zeevi [22] and Whitt [36] also examined queueing models with decision structures that closely resemble ours.

In this article, our main objective is to analyze the fundamental properties of the stochastic fluid makespan problem. In particular, we view the major contributions of this work as follows: (1) We determine necessary and sufficient conditions for such problems to be "well posed" (i.e., to have a finite optimal solution); (2) We give examples of specialized problem structures that are analytically tractable; (3) We provide bounds using convexity results for makespan as a function of the allocation decision and of the stochastic parameters; and, (4) We formulate the makespan problem as a numerically tractable stochastic nonlinear program. In general, stochastic optimization models optimize a performance measure that might be represented by any number of means (e.g., a mathematical program, an analytic queueing model, or a simulation model). A primary contribution of this article is to structurally characterize a performance measure captured by a stochastic fluid model in a way that can then be exploited in solving the associated stochastic program. We hope this leads to further investigation of solution techniques for the problem we pose. Moreover, many systems allow for decision structures that lie between our time-static structure and a real-time control structure. So, we believe that this work provides a stepping stone to the development of related models with more complex, multistage decision structures.

## 2. MODELING AND NOTATION

In this article, we consider a fluid model in which multiple classes of fluid, indexed by $k \in K$, flow through a system consisting of stations indexed by $j \in J$. We envision each class of fluid being stored in a buffer, which we refer to as buffer $k$. We assume $|J| \leq |K|$ and that each class $k$ is served by a unique station $\sigma_k \in J$. On the other hand, station $j$ drains a set of buffers denoted by $\mathcal{C}_j$, where $\mathcal{C}_j = \{k \mid \sigma_k = j\}$. The system starts with an initial inventory $a_k$ in each buffer $k$. Fluid arrives to buffer $k$ from outside the system at rate $\alpha_k$. If station $\sigma_k$ allocates all of its effort to buffer $k$, it takes $m_k$ units of time to drain one unit of fluid from buffer $k$. When subscripts are omitted, $a, \alpha$, and $m$ denote the vector forms of the above parameters. All vectors are assumed to be column vectors. After the fluid leaves buffer $k$, some portion of the fluid is routed to the buffers in the system and the remaining portion leaves the system. The proportion of fluid that is routed to buffer $l$ from buffer $k$ is denoted $p_{kl}$. The $|K| \times |K|$ matrix $P$, with elements $p_{kl}$, is called the routing matrix. In this work, matrices are denoted by uppercase letters; to denote the $k$th row of a matrix, the superscript $k$ notation is used. To avoid confusion, when a superscript is used as a power operator, the matrix is written in parentheses. We use $I$ to denote an appropriately dimensioned identity matrix and $e$ to denote the vector of all 1s.

The system described earlier is called a multiclass fluid network. An example of a multiclass fluid network with two stations and four buffers is given in Figure 1. A multiclass fluid process is given by $(Z(t), T(t))$ for $t \geq 0$. In this notation, $Z(t)$ and $T(t)$ are $|K|$ dimensional. $T_k(t)$ is the total amount of effort (in units of time) spent to drain buffer $k$ up to time $t$, and $Z_k(t)$ gives the amount of fluid in buffer $k$ at time $t$. In the above notation, $Z(0) = a$. We also define the service-time matrix, $M = \text{diag}(m)$. Then the dynamical equations governing the fluid process are, for all $t \geq 0$,

$$Z(t) = a + \alpha t - (I - P')M^{-1}T(t), \tag{1a}$$

$$\sum_{k \in \mathcal{C}_j} T_k(t) \leq t, \qquad j \in J, \tag{1b}$$

$$Z(t) \geq 0, \tag{1c}$$

$$T_k(\cdot) \text{ nondecreasing, } T_k(0) = 0 \text{ for each } k \in K. \tag{1d}$$

A control policy is defined by a set of functions $\{T_k(t), k \in K\}$ on $[0, \infty)$. Once a policy $T(\cdot)$ is specified, then $Z(\cdot)$ is determined by (1a). If the resulting $Z(\cdot)$ satisfies (1c) and $T(\cdot)$ satisfies (1b) and (1d), then the solution is a feasible fluid solution.

Now, suppose we define $v_k(t) = \dot{T}_k(t)$ for all $k$ and all $t \geq 0$ for which the derivative exists. It can be shown that $T(\cdot)$ is absolutely continuous and so its derivatives exist a.e. Then the functions $v_k(\cdot)$ provide an equivalent way to specify the control. One can interpret $v_k(t)$ as the instantaneous percentage of effort at station $\sigma_k$ devoted to draining buffer $k$. In the stochastic setting we will find it easier to specify a control policy via $v(\cdot) \equiv (v_k(\cdot))$.

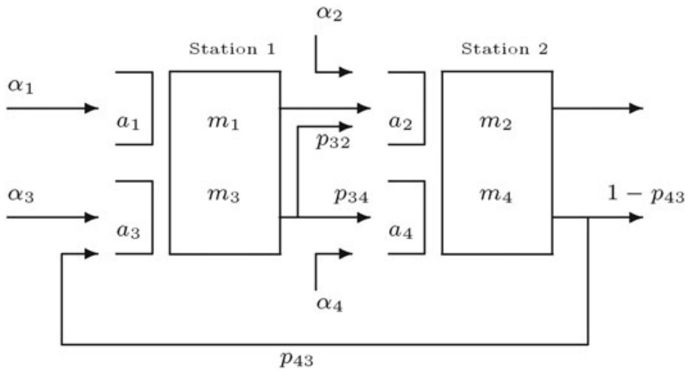

**FIGURE 1.** Multiclass network with two stations and four buffers.

A policy is said to be stationary if these percentage allocations are not functions of time when there is fluid in a buffer. In particular, if buffer $k$ has fluid, $v_k$ percentage of time is dedicated to draining that fluid. When the buffer is empty, we dedicate no more than $v_k$ percentage of time to keep it empty. Clearly, we must then have

$$\sum_{k \in \mathcal{C}_j} v_k \leq 1, \quad \forall j \in J. \tag{2}$$

The makespan of a fluid network is the time that the network is actually drained [i.e., the minimum $t$ such that $Z(t) = 0$]. In this work, we analyze the problem of minimizing the makespan of a given fluid network by deciding on the allocation of effort at each station. Studying the makespan of fluid networks only makes sense if the networks are *open*; that is, all fluid in the system will eventually leave the system. This notion makes more intuitive sense for queueing networks with discrete customers, but it turns out that it is necessary to adopt the same notion for fluid networks. To simplify the notation throughout the article we define

$$Q = I + P' + (P')^2 + \cdots .$$

An open fluid network is one for which the sum above converges. In that case, $Q$ is well defined and its expression reduces to $Q = (I - P')^{-1}$. We refer the reader to [11] for a detailed analysis of open fluid networks.

To ensure that the network can be drained, we need to enforce further conditions on the network parameters. The effective arrival rate of class $k$ is $Q^k\alpha$, and so the amount of work that arrives to the system in unit time, destined for buffer $k$, is given by $Q^k\alpha m_k$. To be able to eventually drain the system, each station $j$ must have enough capacity to process the total work that arrives to the system and is destined for the buffers in $\mathcal{C}_j$; that is, the following inequalities must hold:

$$\sum_{k \in \mathcal{C}_j} Q^k \alpha m_k \leq 1, \quad \forall j \in J. \tag{3}$$

The conditions given in (3) are called *the usual traffic conditions* in the literature. When the inequality holds strictly, we say that *the strict usual traffic conditions hold.*

In [34], the makespan problem is examined for reentrant lines in a deterministic setting; that is, the parameters of the system are known deterministically at the time of decision-making. A reentrant line is a special type of multiclass queueing network, where only the first buffer receives exogenous input and proportional routing is not allowed in the system. In [34] it is shown that, if the strict usual traffic conditions hold, there exists a policy that drains a reentrant line in finite time. A similar result holds for multiclass fluid networks, as shown in [10].

The main focus of this work is how to minimize the expected makespan of a multiclass fluid system over stationary policies, when the parameters $a, \alpha$, and $m$ are not known deterministically at the time of decision-making. The stationarity assumption is restrictive, since it can be shown that, in some cases, all stationary policies are suboptimal when we allow inclusion of time-varying policies. It would certainly be of interest in subsequent research to explore the more complex case with this restriction removed. We denote the sample space of the random variables by $\Omega$ and denote a sample point in the sample space by $\omega \in \Omega$. A random variable $x$ is generally denoted by $\tilde{x}$, and a realization of this random variable under $\omega \in \Omega$ is denoted by $x^{\omega}$.

## 3. THE MAKESPAN PROBLEM

Under a given policy, the makespan of a fluid network is defined as the time that the system reaches the empty state. We seek a policy that drains the system in minimal time. In the deterministic version of the problem, we assume that the parameters of the system (i.e., $a, \alpha$, and $m$) are known deterministically at the time of decision-making. This problem can be formulated as follows:

$$t^* = \min \int_0^{\infty} 1_{\{e'Z(t)>0\}}\, dt \quad \text{s.t. (1a)–(1d).}$$

Here, $1_{\{e'Z(t)>0\}}$ is the indicator function, which takes value 1 if there is fluid in the network at time $t$ and value 0 otherwise. Taking the integral over time, we obtain the total time that there is fluid in the system. Under a stationary policy, the system stays empty after it is drained; hence, the value of the integral is the makespan of our system.

The solution to the deterministic problem of minimizing makespan for general multiclass fluid networks is given in [10, Chap. 12] (a solution for the special case of a fluid reentrant line is given in [34]). To better understand this result and our subsequent developments, we start by calculating the total workload in the buffers.

The amount that should be emptied from the buffers until all buffers are drained can be written in two parts. The first part is the amount of fluid that flows from the buffers due to the initial fluid inventory and is given by

$$a + P'a + (P')^2 a + (P')^3 a + \cdots = Qa.$$

The second part is the amount of fluid that arrives to the system exogenously up to time $t$ and is given by

$$\alpha t + P'\alpha t + (P')^2 \alpha t + \cdots = Q\alpha t, \quad \forall t \in [0, \infty).$$

Using the above calculations, we can compute the total cumulative workload for buffer $k$ up to time $t$ as

$$(Q^k a + Q^k \alpha t) m_k. \tag{4}$$

We are now prepared to present the solution to the deterministic makespan problem. From the above calculations one can see that for each $j$ the numerator on the right-hand side of Eq. (5) below is the total workload due to initial fluid inventory and the denominator represents the percentage of time available for processing initial fluid, assuming the workload due to incoming fluid is processed immediately.

Theorem 3.1: *If the usual traffic conditions for a multiclass fluid network hold, then a lower bound for the makespan is*

$$t^* \geq t^{LB} = \max_{j \in J} \left\{ \frac{\sum_{k \in C_j} Q^k a m_k}{1 - \sum_{k \in C_j} Q^k \alpha m_k} \right\}, \tag{5}$$

*and this value can be attained. Conversely, if the usual traffic conditions are violated, then the makespan is infinite for every policy.*

Proof: See Chen and Yao [10, p. 384]. ■

If a numerator in (5) is zero (i.e., there is no initial fluid to be processed at $j$), then the associated ratio is zero, regardless of the value of the denominator. Otherwise, if the usual traffic conditions hold but they do not hold strictly, then the lower bound (5) is infinite. In this case, we regard the lower bound as being attained, as stated in Theorem 3.1, since the makespan is also infinite.

So far, we have assumed that we know the parameters of the system deterministically at the time we select our control. In this case, Theorem 3.1 as given in [10] provides a simple closed-form expression for an optimal policy. Furthermore, that policy, for the deterministic problem, is a stationary policy. However, in this work our main focus is on the makespan problem where the parameters $a, \alpha$, and $m$ are only known via a probability distribution, and in this case we restrict the control a priori to stationary policies for the reasons described in Section 1.

For the stochastic makespan problem, the time at which the system drains is a random variable, so a natural objective is to minimize the expected value of the makespan. Perhaps the simplest approach to the stochastic makespan problem is to attempt to solve the problem as in the deterministic case, using the expected values of the stochastic parameters; that is, we use the solution for the deterministic makespan problem, where $a$, $\alpha$, and $m$ are replaced by their population means. This solution is called the expected value solution. Despite being commonly used in practice, the expected value solution can be drastically suboptimal when one of the parameter vectors $a, \alpha$, or $m$ is random. We now show that the expected value solution may lead to an infinite expected makespan, even if there are feasible solutions where the expected makespan is finite.

Consider the network with one station and two buffers shown in Figure 2. Let $a = (0, 6)$, $m = (1, 1)$, and $\alpha$ be random with $\mathbb{P}(\tilde{\alpha} = (0, 0)) = \mathbb{P}(\tilde{\alpha} = (1/4, 1/4)) = 1/2$. As a result, $v_{EV} = (1/8, 7/8)$ is the expected value solution; that is, the effort allocated for buffer 1 is just enough to serve the expected inflow and the rest is devoted to drain buffer 2. It is easy to see that with probability 1/2, this solution does not drain the system (i.e., when the scenario $\tilde{\alpha} = (1/4, 1/4)$ occurs). Hence, the expected makespan is infinite if $v_{EV}$ is employed. However, the solution $v^* = (1/4, 3/4)$ yields

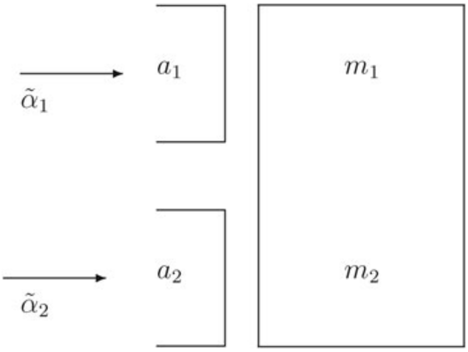

**FIGURE 2.** Network with one station and two buffers.

an expected makespan of 10. It is possible to find a similar example when only $m$ is random. For the case in which $a$ is random, the expected value solution yields a finite makespan with probability 1 if there is a feasible solution with finite expected makespan. Nevertheless, even in this case it is possible to find examples in which the expected value solution is suboptimal. For a detailed discussion on the suboptimality of the expected value solution, we refer the reader to [9].

In general, the stochastic makespan problem cannot be solved using expected values, so we instead attack the problem using stochastic programming techniques. We have already introduced one set of structural constraints for our problem, namely (2). The next step in formulating the problem is to mathematically represent the makespan in terms of $v$ and the random parameters ($\tilde{a}$, $\tilde{\alpha}$, $\tilde{m}$). To do so, we use the fact that if the total workload of buffer $k$ is to be drained at time $t_k$, the associated effort expended must equal the amount of work that has arrived to buffer $k$ up to time $t_k$; that is,

$$v_k t_k = Q^k \tilde{a}\tilde{m}_k + Q^k \tilde{\alpha}\tilde{m}_k t_k, \quad \forall k \in K.$$

Solving this equation for $t_k$ we obtain

$$t_k = \frac{Q^k \tilde{a}\tilde{m}_k}{v_k - Q^k \tilde{\alpha}\tilde{m}_k}, \quad \forall k \in K. \tag{6}$$

To interpret the expression given in (6), observe that the denominator gives the remaining percentage of effort available, if all the work for buffer $k$ due to exogenous arrivals is removed from the system as soon as it arrives. The $t_k$ given in (6) is the time at which buffer $k$ has processed all of the work initially present in the system, assuming it drains the workload due to exogenous arrivals as soon as it arrives. Therefore, the makespan of the system under allocation $v$ is the time when all buffers empty their initial amounts from the system, and it is given by

$$\mathrm{MS}(v,\tilde{a},\tilde{\alpha},\tilde{m}) = \begin{cases} \max_{k\in K}\left\{\dfrac{Q^k\tilde{a}\tilde{m}_k}{v_k - Q^k\tilde{\alpha}\tilde{m}_k}\right\}, & v_k > Q^k\tilde{\alpha}\tilde{m}_k, \forall k \in K \\ \infty, & \text{otherwise.} \end{cases} \tag{7}$$

A necessary condition to have a finite expected makespan under a stationary policy is

$$v_k \geq Q^k\tilde{\alpha}\tilde{m}_k, \quad \forall k \in K, \ \text{w.p.1.} \tag{8}$$

The inclusive inequality in constraint (8) allows for the possibility of infinite makespan, but we minimize with respect to decision vector $v$ and so the expected makespan will be finite whenever possible in the optimization model. Additionally, note that nonnegativity of the allocation decisions, $v \geq 0$, is ensured by (8).

Summarizing the development in (2), (7), and (8), we obtain the following formulation for the stochastic makespan problem:

$$\min_{v} \quad \mathbb{E}\left(\max_{k\in K}\left\{\frac{Q^k\tilde{a}\tilde{m}_k}{v_k - Q^k\tilde{\alpha}\tilde{m}_k}\right\}\right) \tag{9a}$$

$$\text{s.t.} \quad \sum_{k\in\mathcal{C}_j} v_k \leq 1, \ \forall j \in J \tag{9b}$$

$$v_k \geq Q^k\tilde{\alpha}\tilde{m}_k, \ \forall k \in K, \ \text{w.p.1.} \tag{9c}$$

The above formulation yields insight to the deterministic makespan problem; that is,

$$\min_{v} \quad \max_{k\in K}\left\{\frac{Q^k a m_k}{v_k - Q^k\alpha m_k}\right\} \tag{10a}$$

$$\text{s.t.} \quad \sum_{k\in\mathcal{C}_j} v_k \leq 1, \ \forall j \in J \tag{10b}$$

$$v_k \geq Q^k\alpha m_k, \ \forall k \in K. \tag{10c}$$

Specifically, we now show that the deterministic makespan problem is *separable* by station.

THEOREM 3.2: *If the strict usual traffic conditions hold, then the deterministic makespan problem* (10) *can be solved by solving the station makespan problems separately. Specifically, let*

$$\bar{v}^j \in \arg\min_{V^j}\max_{k\in\mathcal{C}_j}\left\{\frac{Q^k a m_k}{v_k - Q^k\alpha m_k}\right\}, \ j \in J, \tag{11}$$

*where* $V^j = \{[v_k]_{k\in\mathcal{C}_j} : \sum_{k\in\mathcal{C}_j} v_k \leq 1, v_k \geq Q^k\alpha m_k, k \in \mathcal{C}_j\}$. *Then,* $v^* = [\bar{v}^j]_{j\in J}$ *solves* (10).

Proof: Constraints (10b) and (10c) are equivalent to $[v^j]_{j\in J} \in \times_{j\in J} V^j$. The usual traffic conditions (3) ensure $V^j \neq \emptyset$, $\forall j \in J$ and, hence, that (10) is feasible. Strictness of these conditions ensures a finite makespan. Formulation (10) can be written as

$$\begin{aligned}\min_{[v^j\in V^j]_{j\in J}} \max_{k\in K} \left\{ \frac{Q^k a m_k}{v_k - Q^k \alpha m_k} \right\} &= \min_{[v^j\in V^j]_{j\in J}} \max_{j\in J} \max_{k\in \mathcal{C}_j} \left\{ \frac{Q^k a m_k}{v_k - Q^k \alpha m_k} \right\} \\ &= \max_{j\in J} \min_{v^j\in V^j} \max_{k\in \mathcal{C}_j} \left\{ \frac{Q^k a m_k}{v_k - Q^k \alpha m_k} \right\}. \end{aligned} \qquad \textbf{(12)}$$

The inner minimization in (12) is equivalent to that in (11), and the proof is complete. ■

Theorem 3.2 shows that even if the output of one station provides input to another station, the optimal allocations of effort can be determined separately. Theorem 3.2 does not extend in general to the stochastic makespan problem. For stochastic problems, the following example shows that this separation result fails to hold even when there is no fluid flow between the stations.

Consider the two-station three-buffer network in Figure 3, where there is no input, only the initial inventory is random, and there is no flow between buffers. The service times are as given in Figure 3, and let $\mathbb{P}(\tilde{a} = (5, 1, 100)) = \mathbb{P}(\tilde{a} = (1, 5, 0)) = 1/2$. Obviously, for the second station, we allocate $v_3 = 1$. If we solve the stochastic problem for station 1 without considering the second station, we obtain $v_1 = v_2 = 1/2$, which leads to an expected network makespan of 55. However, setting $v = (1/6,\ 5/6,\ 1)$ yields an expected makespan of 53.

In the above example, the suboptimality of the allocation based on optimizing the stations separately arises as a result of dependence in the random vector governing the initial inventory. Next, suppose that $\mathbb{P}(\tilde{a} = (10, 0, 14)) = \mathbb{P}(\tilde{a} = (0, 4, 14)) = 1/2$. In this problem, $a_3$ is deterministic and therefore independent of $(a_1, a_2)$. When the problem is solved for the stations separately, we obtain the optimal allocation as $v = (0.613, 0.387, 1)$, which yields an expected makespan of 15.15. However, when the problem is solved taking both stations into account, the optimal allocation is

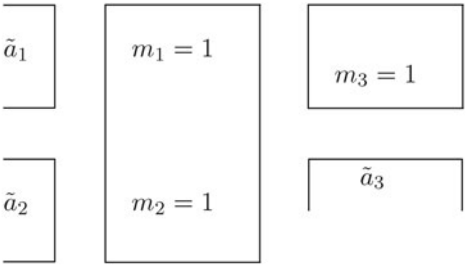

Figure 3. Nonseparable stochastic network.

$v^* = (0.714, 0.286, 1)$, yielding a makespan of 14. Therefore, the separability property does not necessarily hold, even when the stations are completely independent.

Equation (12) in the proof of Theorem 3.2 follows from exchanging the order of minimization and maximization. However, when we have stochastic parameters, the analog of (12) is

$$\min_{[v^j \in V^j]_{j\in J}} \mathbb{E}\left(\max_{k\in K}\left\{\frac{Q^k a m_k}{v_k - Q^k \alpha m_k}\right\}\right) = \min_{[v^j \in V^j]_{j\in J}} \mathbb{E}\left(\max_{j\in J}\max_{k\in \mathcal{C}_j}\left\{\frac{Q^k a m_k}{v_k - Q^k \alpha m_k}\right\}\right).$$

With the presence of the expectation operator, it is no longer possible to interchange the optimization operations. As a result, Theorem 3.2 does not hold for the stochastic case.

## 4. EXISTENCE OF A FINITE OPTIMAL SOLUTION

For the deterministic makespan problem, Theorem 3.1 implies that if the usual traffic conditions are strictly satisfied, a solution that yields a finite makespan exists. The natural question that then arises in the stochastic problem is whether there is a solution that yields a finite expected makespan if the usual traffic conditions hold with probability 1. Unfortunately, it turns out that the almost sure usual traffic conditions do not, in general, guarantee a finite expected makespan in the stochastic makespan problem. As an example consider the network in Figure 2 when $\alpha$ is random. Suppose $a$ and $m$ are deterministic with $a = (0,0)$ and $m = (1,1)$, and let $\mathbb{P}(\tilde{\alpha} = (2/3, 0)) = \mathbb{P}(\tilde{\alpha} = (0, 2/3)) = 1/2$. The traffic intensity, $\sum_k \tilde{\alpha}_k m_k$, is $2/3$ under both scenarios. So, the usual traffic conditions are satisfied in both scenarios. However, since $\tilde{\alpha}_1$ takes value $2/3$ in the first scenario, it is necessary to have $v_1 \geq 2/3$ for a finite expected makespan. By a symmetric argument, we also need $v_2 \geq 2/3$ for the second scenario. These conditions on $(v_1, v_2)$ are inconsistent with (9b); that is, there is no effort allocation that yields a finite expected makespan, even though the usual traffic conditions are satisfied for each of the scenarios.

Motivated by the above example, we now derive necessary and sufficient conditions that guarantee the existence of an allocation with finite expected makespan. To do so, we need the notion of the *essential supremum* of a random variable $\tilde{X}$:

$$\operatorname{ess\,sup}\{\tilde{X}\} \equiv \inf\{x : \mathbb{P}(\tilde{X} > x) = 0\}.$$

THEOREM 4.1: *Consider a fluid makespan problem with stochastic parameters, and let $S_k \equiv \operatorname{ess\,sup}\{Q^k \tilde{\alpha} \tilde{m}_k\}$. The necessary and sufficient conditions for a solution with a finite expected makespan to exist are as follows:*

(a) $\sum_{k\in\mathcal{C}_j} S_k \leq 1, \forall j \in J.$

(b)

$$\mathbb{E}\left(\frac{Q^k \tilde{a} \tilde{m}_k}{S_k + \left(1 - \sum_{l\in\mathcal{C}_{\sigma_k}} S_l\right)/|\mathcal{C}_{\sigma_k}| - Q^k \tilde{\alpha} \tilde{m}_k}\right) < \infty, \quad \forall k \in K.$$

Proof: If conditions (a) and (b) hold, then allocation vector $v$, where

$$v_k = S_k + \frac{1 - \sum_{l \in \mathcal{C}_{\sigma_k}} S_l}{|\mathcal{C}_{\sigma_k}|}$$

for all $k \in K$, satisfies (9b) and (9c). From condition (b) we know that

$$\mathbb{E}\left(\frac{Q^k \tilde{a} \tilde{m}_k}{v_k - Q^k \tilde{\alpha} \tilde{m}_k}\right) < \infty, \quad \forall k \in K.$$

The buffer index set $K$ is finite and hence so is the objective function (9a) evaluated at this $v$. This proves the sufficiency of conditions (a) and (b) for the existence of a solution with a finite expected makespan.

Now, suppose that (a) does not hold; that is, there is a station $j^* \in J$ with $\sum_{k \in \mathcal{C}_{j^*}} S_k > 1$. Constraint (9c) is equivalent to $v_k \geq S_k$, for all $k \in K$. Thus, (9b) for $j = j^*$ and (9c) for $k \in \mathcal{C}_{j^*}$ are inconsistent; that is, (9) is infeasible and hence has no solution with finite expected makespan.

Finally, suppose that (b) does not hold; that is,

$$\mathbb{E}\left(\frac{Q^{k^*} \tilde{a} \tilde{m}_{k^*}}{S_{k^*} + (1 - \sum_{l \in \mathcal{C}_{\sigma_{k^*}}} S_l)/|\mathcal{C}_{\sigma_{k^*}}| - Q^{k^*} \tilde{\alpha} \tilde{m}_{k^*}}\right) = \infty \qquad \textbf{(13)}$$

for some $k^* \in K$. If the system defined by (9b) and (9c) is infeasible, then the proof is complete, so we restrict attention to the case when $\sum_{l \in \mathcal{C}_{\sigma_{k^*}}} S_l \leq 1$ and, in turn, consider the cases when this inequality holds with equality and is strict.

*Case 1*: $\sum_{l \in \mathcal{C}_{\sigma_{k^*}}} S_l = 1$. In this case, in all feasible allocations, $v_{k^*} = S_{k^*}$. By (13),

$$\mathbb{E}\left(\frac{Q^{k^*} \tilde{a} \tilde{m}_{k^*}}{S_{k^*} - Q^{k^*} \tilde{\alpha} \tilde{m}_{k^*}}\right) = \infty,$$

and hence the objective function (9a) is also infinite for any feasible $v$.

*Case 2*: $\sum_{l \in \mathcal{C}_{\sigma_{k^*}}} S_l < 1$. So, there exists an $\epsilon > 0$ such that $\sum_{l \in \mathcal{C}_{\sigma_{k^*}}} S_l + \epsilon = 1$. Thus,

$$\mathbb{E}\left(\frac{Q^{k^*} \tilde{a} \tilde{m}_{k^*}}{S_{k^*} + \epsilon/|\mathcal{C}_{\sigma_{k^*}}| - Q^{k^*} \tilde{\alpha} \tilde{m}_{k^*}}\right) \leq \frac{|\mathcal{C}_{\sigma_{k^*}}| \mathbb{E}\left(Q^{k^*} \tilde{a} \tilde{m}_{k^*}\right)}{\epsilon},$$

which implies $\mathbb{E}\left(Q^{k^*} \tilde{a} \tilde{m}_{k^*}\right) = \infty$. For any feasible allocation $v$,

$$\mathbb{E}\left(\frac{Q^{k^*} \tilde{a} \tilde{m}_{k^*}}{v_{k^*} - Q^{k^*} \tilde{\alpha} \tilde{m}_{k^*}}\right) \geq \mathbb{E}\left(\frac{Q^{k^*} \tilde{a} \tilde{m}_{k^*}}{1}\right) = \infty.$$

■

When conditions (a) and (b) of Theorem 4.1 hold, the proof of the theorem specifies a feasible solution that yields a finite expected makespan. We assume the existence of a finite optimal solution for the remainder of this article.

## 5. SPECIAL CASES

In Section 3, we discussed examples showing that the solution obtained by using the expected values of the random parameters need not be optimal for the stochastic makespan problem. In general, it is not possible to state the solution of the stochastic problem analytically. That said, the purpose of this section is to describe special cases in which it is possible to characterize the solution analytically. We explore the special cases that arise when only one set of the parameters $a$, $\alpha$, or $m$ is random.

Although the expected value solution does not, in general, solve the stochastic problem, we can ask: If we have a stochastic fluid system, in which one of the stations is "deterministic," is it possible to say anything about the solution? The following theorem answers this question.

THEOREM 5.1: *Assume model* (9) *has a solution with finite expected makespan. If* $j^* \in J$ *satisfies* $Q^k a^{\omega_1} m_k^{\omega_1} = Q^k a^{\omega_2} m_k^{\omega_2} \equiv \beta_k$ *and* $Q^k \alpha^{\omega_1} m_k^{\omega_1} = Q^k \alpha^{\omega_2} m_k^{\omega_2} \equiv \rho_k$, $\forall \omega_1, \omega_2 \in \Omega$ *and* $\forall k \in \mathcal{C}_{j^*}$, *then there is an optimal solution,* $v^*$, *to the stochastic makespan problem with*

$$v_k^* = \frac{\beta_k - \beta_k \sum_{l \in \mathcal{C}_{j^*}} \rho_l + \sum_{l \in \mathcal{C}_{j^*}} \beta_l \rho_k}{\sum_{l \in \mathcal{C}_{j^*}} \beta_l}, \quad \forall k \in \mathcal{C}_{j^*}. \tag{14}$$

PROOF: We first note that $v_k^*$, $k \in \mathcal{C}_{j^*}$, satisfies (9c) because the numerator of (14) is positive by condition (a) of Theorem 4.1, and (9b) for $j = j^*$ holds since

$$\sum_{k \in \mathcal{C}_{j^*}} v_k^* = \frac{\sum_{k \in \mathcal{C}_{j^*}} \beta_k - \left(\sum_{k \in \mathcal{C}_{j^*}} \beta_k\right)\left(\sum_{l \in \mathcal{C}_{j^*}} \rho_l\right) + \left(\sum_{l \in \mathcal{C}_{j^*}} \beta_l\right)\left(\sum_{k \in \mathcal{C}_{j^*}} \rho_k\right)}{\sum_{l \in \mathcal{C}_{j^*}} \beta_l} = 1.$$

Let $h_k(v_k)$ denote the draining time of buffer $k \in \mathcal{C}_{j^*}$ [i.e., the right-hand side of (6)]. The draining time of the last buffer at station $j^*$ is $\max_{k \in \mathcal{C}_{j^*}} h_k(v_k)$. Note that $h_k$ is a decreasing function over feasible allocations and $h_k(v_k^*) = \sum_{l \in \mathcal{C}_{j^*}} \beta_l / (1 - \sum_{l \in \mathcal{C}_{j^*}} \rho_l)$; that is, the draining time is equal for all buffers $k \in \mathcal{C}_{j^*}$. This coupled with $\sum_{k \in \mathcal{C}_{j^*}} v_k^* = 1$ implies

$$\max_{k \in \mathcal{C}_{j^*}} \left\{h_k(v_k^*)\right\} \leq \max_{k \in \mathcal{C}_{j^*}} \{h_k(v_k)\} \tag{15}$$

for all feasible allocations $v_k, k \in \mathcal{C}_{j^*}$. Suppose $v^{**}$ solves the stochastic makespan problem and extend $v_k^*$ from (14) to $v_k^* = v_k^{**}, k \in K \setminus \mathcal{C}_{j^*}$. From (15), we have

$$\max_{k \in \mathcal{C}_{j^*}} \left\{ \frac{Q^k a^\omega m_k^\omega}{v_k^* - Q^k \alpha^\omega m_k^\omega} \right\} \leq \max_{k \in \mathcal{C}_{j^*}} \left\{ \frac{Q^k a^\omega m_k^\omega}{v_k^{**} - Q^k \alpha^\omega m_k^\omega} \right\}, \quad \forall \omega \in \Omega. \tag{16}$$

We know that $v^*$ and $v^{**}$ drain all other buffers at the same time; hence,

$$\max_{k\in K}\left\{\frac{Q^k a^\omega m_k^\omega}{v_k^* - Q^k \alpha^\omega m_k^\omega}\right\} \le \max_{k\in K}\left\{\frac{Q^k a^\omega m_k^\omega}{v_k^{**} - Q^k \alpha^\omega m_k^\omega}\right\}, \quad \forall\omega\in\Omega.$$

Taking expectations,

$$\mathbb{E}\left(\max_{k\in K}\left\{\frac{Q^k\tilde{a}\tilde{m}_k}{v_k^* - Q^k\tilde{\alpha}\tilde{m}_k}\right\}\right) \le \mathbb{E}\left(\max_{k\in K}\left\{\frac{Q^k\tilde{a}\tilde{m}_k}{v_k^{**} - Q^k\tilde{\alpha}\tilde{m}_k}\right\}\right).$$

Hence, $v^*$ also solves the stochastic makespan problem. ■

Theorem 5.1 implies that if the random parameters defining our stochastic makespan problem have a certain structure, then the expected value solution solves the stochastic problem. In the next three subsections, we clarify this implication by examining three special, intuitive cases.

### 5.1. Random Incoming Rates

In this subsection, we assume that only the incoming rate vector $\alpha$ is random and that it has a special probabilistic structure. Specifically, we assume that randomness is observed proportionally for all buffers; that is, there is a deterministic base rate vector $\alpha^0$, and for any scenario, $\omega\in\Omega$, the rate vector can be represented as $N^\omega\alpha^0$. Here, $N^\omega$ is a scalar determined by scenario $\omega$. This is equivalent to assuming that fluid arrives to the system from a single source with an unknown rate, but it is distributed to the stations in the system according to fixed proportions. Note that since fluid reentrant lines have exogenous arrivals to only one buffer, this structural assumption always holds for stochastic makespan problems in such networks.

With the above assumption, we can construct the following special case.

THEOREM 5.2: *Assume model* (9) *has a solution with finite expected makespan. If $\tilde{\alpha} = \tilde{N}\alpha^0$ and $j^*\in J$ satisfies $Q^k\alpha^0/Q^k a = Q^l\alpha^0/Q^l a, \forall k,l\in\mathcal{C}_{j^*}$, then there is an optimal solution, $v^*$, to the stochastic makespan problem with*

$$v_k^* = \begin{cases} \dfrac{Q^k\alpha^0 m_k}{\sum_{l\in\mathcal{C}_{j^*}} Q^l\alpha^0 m_l} & \text{if } \sum_{l\in\mathcal{C}_{j^*}} Q^l\alpha^0 m_l > 0 \\[2ex] \dfrac{Q^k a m_k}{\sum_{l\in\mathcal{C}_{j^*}} Q^l a m_l} & \text{if } \sum_{l\in\mathcal{C}_{j^*}} Q^l\alpha^0 m_l = 0 \end{cases}, \quad \forall k\in\mathcal{C}_{j^*}. \tag{17}$$

PROOF: We first show that all the buffers in station $j^*$ are drained at the same time for each scenario. If $\sum_{l\in\mathcal{C}_{j^*}} Q^l\alpha^0 m_l = 0$, then $j^*$ satisfies the conditions of Theorem 5.1. Moreover, $v_k^*$ from (17) is identical to that of (14), and hence, from the proof of Theorem 5.1 all of the buffers at $j^*$ are drained at the same time. If $\sum_{l\in\mathcal{C}_{j^*}} Q^l\alpha^0 m_l > 0$, then $\sum_{k \in \mathcal{C}_{j^*}} v_k^* = 1$ and, as earlier, (9c) holds by condition (a) of Theorem 4.1. Then, for any $k, l \in \mathcal{C}_{j^*}$ and $\omega \in \Omega$,

$$\frac{Q^k a m_k}{Q^k \alpha^0 m_k / \sum_{i \in \mathcal{C}_{j^*}} Q^i \alpha^0 m_i - N^\omega Q^k \alpha^0 m_k} = \frac{Q^k a m_k}{Q^k \alpha^0 m_k \left(1/\sum_{i \in \mathcal{C}_{j^*}} Q^i \alpha^0 m_i - N^\omega\right)}$$
$$= \frac{Q^l a m_l}{Q^l \alpha^0 m_l \left(1/\sum_{i \in \mathcal{C}_{j^*}} Q^i \alpha^0 m_i - N^\omega\right)}$$
$$= \frac{Q^l a m_l}{Q^l \alpha^0 m_l / \sum_{i \in \mathcal{C}_{j^*}} Q^i \alpha^0 m_i - N^\omega Q^l \alpha^0 m_l}.$$

Hence, in each scenario, the proposed solution drains all of the buffers at $j^*$ at the same time.

Next, we show that $(v_k^*)_{k \in \mathcal{C}_{j^*}}$ leads to a finite expected draining time for all buffers at $j^*$. By hypothesis, conditions (a) and (b) of Theorem 4.1 are satisfied. Using (a), we know that $\sum_{k \in \mathcal{C}_{j^*}} \operatorname{ess\,sup}\{\tilde{N}\} Q^k \alpha^0 m_k \leq 1$. Hence,

$$\operatorname{ess\,sup}\{\tilde{N}\} \leq 1/ \sum_{k \in C_{j^*}} Q^k \alpha^0 m_k. \tag{18}$$

If the inequality holds strictly, there exists an $\epsilon > 0$ such that

$$\mathbb{E}\left(\frac{Q^k a m_k}{v_k^* - \tilde{N} Q^k \alpha^0 m_k}\right) < \frac{Q^k a m_k}{\epsilon}, \quad \forall k \in \mathcal{C}_{j^*}.$$

On the other hand, if the inequality holds as an equality, then $v_k^* = S_k, \forall k \in \mathcal{C}_{j^*}$. Using condition (b) of Theorem 4.1, we conclude that $v^*$ leads to a finite expected draining time for all buffers at $j^*$.

Suppose $v^{**}$ solves the stochastic makespan problem and extend the definition of $v_k^*$ from (17) to $v_k^* = v_k^{**}, k \in K \setminus \mathcal{C}_{j^*}$. The proof can now be completed using the same argument as in the proof of Theorem 5.1. ■

Note that Theorem 5.2 holds even if there is a $j^*$, such that $Q^k \alpha^0 / Q^k a = Q^l \alpha^0 / Q^l a = \infty \ \forall k, l \in \mathcal{C}_{j^*}$, (i.e., $Q^k a = Q^l a = 0$). Since the necessary and sufficiency conditions are satisfied, (18) implies $Q^k \alpha^0 m_k / \sum_{i \in \mathcal{C}_{j^*}} Q^i \alpha^0 m_i - N^\omega Q^k \alpha^0 m_k \geq 0, \forall \omega \in \Omega$. Hence, the station stays empty for all scenarios and all the buffers are still drained at the same time, so the result follows.

## 5.2. Random Service Rates

In the previous subsection, the arrival rates for all buffers in the system were perfectly correlated. Since the arrival rates might be determined by the same causes in the exogenous environment and there are systems like reentrant lines, such a dependence assumption could naturally arise. However, assuming a similar structure for the system's service rates might be overly restrictive. Fortunately, in the case where service rates are random, similar results hold with a relaxed version of the dependence assumption. In particular, we need only assume that service rates for buffers within the same station are proportional for all scenarios $\omega \in \Omega$; that is, there is a base service time $m^0$, and for any scenario $\omega$, $m_k^\omega = N_j^\omega m_k^0$. Here, $N_j^\omega$ is determined by station $j$ and scenario $\omega$ and can differ by station under the same scenario. This probabilistic structure could arise as follows. Suppose that there are several identical machines at each station with deterministically known service times but that the number of machines in working condition is unknown when the allocation policy must be specified. In this case, a fluid model with the above random service rate structure can serve as a reasonable approximation.

The next theorem allows us to present a result useful for systems in which $a$ and $\alpha$ are not random, and the service rates are correlated in the manner just discussed earlier.

THEOREM 5.3: *Assume model* (9) *has a solution with finite expected makespan. If $j^* \in J$ satisfies $\tilde{m}_k = \tilde{N}_{j^*} m_k^0, k \in \mathcal{C}_{j^*}$ and $Q^k\alpha/Q^k a = Q^l\alpha/Q^l a$, where $k, l \in \mathcal{C}_{j^*}$, then there is an optimal solution, $v^*$, to the stochastic makespan problem with*

$$v_k^* = \begin{cases} \dfrac{Q^k\alpha m_k^0}{\sum_{l\in\mathcal{C}_{j^*}} Q^l\alpha m_l^0} & \text{if } \sum_{l\in\mathcal{C}_{j^*}} Q^l\alpha m_l^0 > 0 \\ \dfrac{Q^k a m_k^0}{\sum_{l\in\mathcal{C}_{j^*}} Q^l a m_l^0} & \text{if } \sum_{l\in\mathcal{C}_{j^*}} Q^l\alpha m_l^0 = 0 \end{cases}, \quad \forall k \in \mathcal{C}_{j^*}. \tag{19}$$

PROOF: Using the same approach as in Theorem 5.2, it can be shown that all buffers at station $j^*$ are drained at the same time for each scenario. Then the result follows from the argument used in the proofs of Theorems 5.1 and 5.2. ■

### 5.3. Random Initial Inventory

As a final special case, we consider a system in which $\alpha$ and $m$ are deterministic but the initial inventory vector $\tilde{a}$ is random with perfectly correlated components.

THEOREM 5.4: *Assume model* (9) *has a solution with finite expected makespan. If $\tilde{a} = \tilde{N}a^0$, then*

$$v_k^* = \frac{\beta_k - \beta_k \sum_{l\in\mathcal{C}_{\sigma_k}} \rho_l + \sum_{l\in\mathcal{C}_{\sigma_k}} \beta_l \rho_k}{\sum_{l\in\mathcal{C}_{\sigma_k}} \beta_l}, \quad \forall k \in K, \tag{20}$$

*where $\beta_k = Q^k a^0 m_k$ and $\rho_k = Q^k\alpha m_k$, solves both the expected value and stochastic makespan problems.*

PROOF: Notice that,

$$\mathbb{E}\left(\max_{k\in K}\left\{\frac{Q^k\tilde{N}a^0m_k}{v_k - Q^k\alpha m_k}\right\}\right) = \mathbb{E}(\tilde{N})\max_{k\in K}\left\{\frac{Q^k a^0 m_k}{v_k - Q^k\alpha m_k}\right\}.$$

Hence, minimizing $\max_{k\in K}\left\{Q^k a^0 m_k/(v_k - Q^k\alpha m_k)\right\}$ subject to (9b) and (9c) yields an allocation that solves both the stochastic and expected value versions of the makespan problem. The form of $v^*$ given in (20) then follows by applying Theorem 5.1. ■

## 6. SOLUTION METHODS

In Section 5 we presented a number of cases in which the stochastic makespan problem can be solved analytically, specifically by using the so-called expected value solution. However, these special cases are of somewhat limited scope. So, in this section we outline methods for solving, or approximately solving, the stochastic makespan problem (9), that are more generally applicable. Our goal here is to give an overview of available solution approaches, depending on the nature of the distribution of $\tilde{\xi} = (\tilde{a}, \tilde{\alpha}, \tilde{m})$, not to carry out a detailed computational study. However, we do provide results that suggest that our optimization model (9) is numerically tractable on moderate- to large-sized networks. (Our test problems have up to 75 buffers and 15 stations.)

Let $h_k(v_k, \xi)$ denote the draining time of buffer $k$ with allocation $v_k$ and parameter realization $\xi = (a, \alpha, m)$. First, we note that for fixed $\xi$, $h_k(v_k, \xi) = Q^k a m_k/(v_k - Q^k\alpha m_k)$ is convex in feasible $v_k$, and hence, so are $\text{MS}(v, \xi) = \max_{k\in K} h_k(v_k, \xi)$ and $\mathbb{E}(\text{MS}(v, \tilde{\xi}))$. This, coupled with the fact that (9c) can be replaced by $v_k \geq S_k \equiv \text{ess sup}\{Q^k\tilde{\alpha}\tilde{m}_k\}$, $\forall k \in K$, means that (9) is a convex optimization problem. If $\Omega$ is finite with a modest number of sample points and with probability mass function $p^\omega = \mathbb{P}(\tilde{\xi} = \xi^\omega)$, $\omega \in \Omega$, then we can solve (9) using a convex nonlinear programming algorithm. One such algorithm is a variant of Kelley's cutting-plane method [24] that handles the nondifferentiability of our objective function that arises from the "$\max_{k\in K}$." In stochastic programming, this algorithm is known as the L-shaped method [33]. The algorithm iteratively solves a master program whose size depends on the dimension of $v$ and evaluates $\mathbb{E}(\text{MS}(v, \tilde{\xi}))$ and its (sub)gradient at the master program solution. The algorithm scales well with $|\Omega|$ because these latter computations separate for each $\omega \in \Omega$ and, hence, can be done quickly.

If $\tilde{\xi}$ has too many (possibly an infinite number of) realizations, we cannot solve (9) exactly, but approximation techniques can be employed. We discuss two approximations: one based on Monte Carlo sampling and the other on deterministically valid bounds.

The Monte Carlo sampling approximation, in its simplest form, entails generating independent and identically distributed (i.i.d.) observations $\tilde{\xi}^1, \ldots, \tilde{\xi}^n$ from the distribution of $\tilde{\xi}$ and solving

$$\min_{v} \quad \frac{1}{n}\sum_{i=1}^{n} \max_{k\in K} h_k(v_k,\tilde{\xi}^i) \tag{21a}$$

$$\text{s.t.} \quad \sum_{k\in C_j} v_k \le 1,\ \forall j \in J \tag{21b}$$

$$v_k \ge S_k,\ \forall k \in K. \tag{21c}$$

Let $v^*(n)$ denote the optimal solution and $z^*(n)$ denote the optimal objective function value of (21). Model (9) has a convex objective function and a compact (and convex) feasible region. As a result, with probability 1, $z^*(n) \to z^*$, where $z^*$ is the optimal value of (9), and all limit points of $\{v^*(n)\}$ solve (9). See the recent review in [31] for these consistency results and other asymptotic properties of the Monte Carlo method. Of course, from the perspective of numerically solving model (21), we can again exploit its structure with application of the cutting-plane method described earlier.

Our second approximation method applies to the special cases of the stochastic makespan problem, when only one set of the stochastic parameters (either $a$, $\alpha$, or $m$) is random or when these three random vectors are independent. This approach uses deterministically valid bounds on the objective function that exploit the convexity of $\mathrm{MS}(v,\cdot)$ with respect to the stochastic parameters.

THEOREM 6.1: *Let $v$ satisfy* (21b) *and* (21c). *Then* $\mathrm{MS}(v,\cdot,\alpha,m)$, $\mathrm{MS}(v,a,\cdot,m)$, *and* $\mathrm{MS}(v,a,\alpha,\cdot)$ *are convex functions on the convex hull of the stochastic parameters' support.*

PROOF: It suffices to show $h_k(v_k,\cdot,\alpha,m)$, $h_k(v_k,a,\cdot,m)$, and $h_k(v_k,a,\alpha,\cdot)$ are convex because, in each case, MS is then the maximum of a finite collection of convex functions and hence is convex.

*Case 1*: $h_k(v_k,\cdot,\alpha,m)$ is a linear function and thus convex.

*Case 2*: $h_k(v_k,a,\cdot,m)$ is the composition of a convex, increasing function, $f(x) = Q^k a m_k/(v_k - x)$, with a linear function and is therefore convex.

*Case 3*: Let $f(m_k) = Q^k a m_k/(v_k - Q^k \alpha m_k)$. The second derivative of $f$ is

$$\frac{d^2 f(m_k)}{dm_k^2} = \frac{Q^k a Q^k \alpha v_k}{(v_k - Q^k \alpha m_k)^3}.$$

Convexity of $f$ and, hence, $h_k(v_k,a,\alpha,\cdot)$ again follow as $v_k$ is feasible. ■

Let $f : \Re^d \to \Re$ be a convex function and $\tilde{\xi}$ be a random $d$-vector. Jensen's inequality provides a well-known lower bound on $\mathbb{E}f(\tilde{\xi})$ [i.e., $\mathbb{E}f(\tilde{\xi}) \ge f(\mathbb{E}\tilde{\xi})$]. When $\tilde{\xi}$

has bounded support, a class of upper bounds on $\mathbb{E}f(\tilde{\xi})$ is provided by the Edmundson–Madansky (EM) inequality. Madansky [26] and Frauendorfer [15] developed this bound in the respective cases when the components of $\tilde{\xi}$ are independent and dependent, assuming that $\tilde{\xi}$'s support is (contained in) a hyperrectangle. These results have been extended to simplicial and general polyhedral domains [13,17]. We represent an EM bound via $\mathbb{E}f(\tilde{\xi}) \leq \mathbb{E}f(\tilde{\xi}_{EM})$, where $\tilde{\xi}_{EM}$ is a random vector taking values only at the extreme points of $\tilde{\xi}$'s support. So, if the domain is a hyperrectangle, computing $\mathbb{E}f(\tilde{\xi}_{EM})$ requires $2^d$ evaluations of $f$ but that number is $d+1$ for a simplicial domain.

Theorem 6.1 allows us to apply the bounds of Jensen and Edmundson–Madansky to the following important special cases of the stochastic makespan problem.

Corollary 6.1: *Let* $\mathrm{MS}(v, \xi)$ *denote the makespan function. If only one set of the stochastic parameters (either a, α, or m) is random or if the subvectors* $\tilde{a}$, $\tilde{\alpha}$, *and* $\tilde{m}$ *of* $\tilde{\xi} = (\tilde{a}, \tilde{\alpha}, \tilde{m})$ *are mutually independent, then*

$$\mathrm{MS}(v, \mathbb{E}\tilde{\xi}) \leq \mathbb{E}\Big(\mathrm{MS}(v, \tilde{\xi})\Big) \leq \mathbb{E}\Big(\mathrm{MS}(v, \tilde{\xi}_{EM})\Big). \qquad \mathbf{(22)}$$

Assuming $\tilde{\xi}$ satisfies the hypothesis of Corollary 6.1, we can solve the makespan problem under the single scenario $\mathbb{E}\tilde{\xi}$ to obtain allocation $v_{EV}$ and optimal value $\underline{z} = \mathrm{MS}(v_{EV}, \mathbb{E}\tilde{\xi}) \leq z^* \equiv \min_v \mathbb{E}(\mathrm{MS}(v, \tilde{\xi}))$. Carrying out this optimization over the feasible region of (21b)–(21c) ensures $v_{EV}$ is feasible to the stochastic makespan problem. Then, we compute $\bar{z} = \mathbb{E}(\mathrm{MS}(v_{EV}, \tilde{\xi}_{EM})) \geq \mathbb{E}(\mathrm{MS}(v_{EV}, \tilde{\xi})) \geq z^*$. If $\bar{z} - \underline{z}$ is sufficiently small, then $v_{EV}$ is a high-quality approximate solution to the stochastic makespan problem. Otherwise, the Jensen and Edmundson–Madansky bounds can be tightened by applying them in conditional fashion to a partition of $\tilde{\xi}$'s support. In this way, the lower and upper bounds of (22) allow us to employ a bounding-and-approximation scheme to (approximately) solve the stochastic makespan problem.

We consider two sets of test problems, each with four networks using buffer–station combinations of 10-5, 25-5, 50-10, and 75-15. Our first set of test problems are reentrant lines. In the reentrant "10-5" test problem, fluid makes two left-to-right passes through the five stations, and in the other three problems, the fluid makes five such passes. In these reentrant problems, all incoming rates are zero except at the first buffer of the first station, and $\tilde{\alpha}_1$ is assumed to be a continuous uniform random variable on $(0, \alpha_1^{\max})$. The second set of test problems are identical to the reentrant lines, except *every* buffer receives exogenous arrivals with each $\tilde{\alpha}_k$ being an independent uniform random variable on $(0, \alpha_k^{\max})$. In all test problems, the parameters $a$ and $m$ are deterministic.

We form the test problems by first randomly selecting values of $m$ and $a$. Then, each such selection forms a *single instance* of a test problem, where $m$ and $a$ are fixed and only $\alpha$ is random. For example, in the 75-15 model, we form a test problem by selecting 75 $m_k$ values uniformly from [0, 1] and we similarly select 75 $a_k$ values from the discrete uniform on $\{1, \ldots, 10\}$. The value of $\alpha_1^{\max}$ in the first set of (reentrant) test problems and the vector $\alpha^{\max}$ in the second set of problems are then selected so that the conditions presented in Theorem 4.1 hold. We note that our simpler reentrant test problems satisfy the first, but not the second, hypothesis of Theorem 5.2. The latter condition fails to hold because initial inventories at the buffers fail to have the required form.

We apply the approximation procedure with deterministic bounds to the reentrant lines in which the randomness is one dimensional, but for reasons alluded to earlier, we do not apply that procedure to the problems with higher-dimensional randomness in which all buffers receive exogenous inflow. The Monte Carlo approximation is applied to both sets of test problems.

The Jensen and Edmundson–Madansky bounds are applied conditionally to a partition of $(0, \alpha_1^{\max})$ with $n = 10{,}000$ equally sized cells for the reentrant test problems. We compute the associated Jensen bound by solving the stochastic makespan problem with 10,000 realizations (i.e., conditional expectations on the 10,000 cells) using the cutting-plane method described earlier. Doing so yields $\underline{z}$ and a solution $v^*(n)$. We then evaluate the Edmundson–Madansky upper bound, $\bar{z}$, at $v^*(n)$, which requires $10,001$ function evaluations of $\mathrm{MS}(v^*(n), \cdot)$. For the four reentrant test problems, the associated percentage gap, $100(\bar{z} - \underline{z})/\underline{z}$ is listed in the second column of Table 1. By increasing the number of cells $n$, we can ensure that the conditional Jensen and EM bounds, as well as the solutions $v^*(n)$, converge to their counterparts for problem (9). Development of this sequential approximation method using the Jensen and EM bounds begins with Huang, Ziemba, and Ben-Tal [21], and adaptive schemes for forming the cell-based partition of the support are described, for example, by [8,14,16,23].

Table 1 also shows the computation time required to solve instances of model (21) for $n = 10,000$ i.i.d. observations of $\tilde{\alpha}_1$ to varying levels of precision, again using the cutting-plane algorithm. All reported CPU times are on a 1.8 GHz, Pentium Xeon dual-processor machine with 1 GB of memory. At each iteration, the algorithm produces upper and lower bounds $\bar{z}^*(n)$ and $\underline{z}^*(n)$ on $z^*(n)$, the optimal value of model (21). The cutting-plane algorithm terminates when $(\bar{z}^*(n) - \underline{z}^*(n))/\underline{z}^*(n) \le \epsilon$. The coefficient of variation of the sample mean objective function of the 75-15 reentrant test problem, with $v = v^*(n)$ and $n = 10,000$, is roughly $10^{-5}$, meaning that there is little point in solving model (21) for more precise values of $\epsilon$. We note that the times to compute the Jensen lower bound with 10,000 cells are essentially the same as those reported in Table 1.

**Table 1.** Gap Between the EM and Jensen Bounds, and Computation Times (s) for Reentrant Lines of Different Sizes for 10,000 Sample Points

| Buffers–Stations | EM–J (%) | $\epsilon = 10^{-1}$ | $\epsilon = 10^{-2}$ | $\epsilon = 10^{-3}$ | $\epsilon = 10^{-4}$ | $\epsilon = 10^{-5}$ | $\epsilon = 10^{-6}$ |
|---|---|---|---|---|---|---|---|
| 10-5 | 0.0000 | 2 | 2 | 3 | 4 | 4 | 5 |
| 25-5 | 0.0420 | 22 | 30 | 60 | 213 | 411 | 518 |
| 50-10 | 0.0004 | 89 | 141 | 334 | 1,469 | 4,263 | 5,331 |
| 75-15 | 0.0046 | 208 | 861 | 1,877 | 3,210 | 6,967 | 11,212 |

**Table 2.** Computation Times (s) for the Monte Carlo Approximation with 10,000 Sample Points Applied to the Test Problems Where Each Buffer Receives Exogenous Arrivals

| Buffers–Stations | $\epsilon = 10^{-1}$ | $\epsilon = 10^{-2}$ | $\epsilon = 10^{-3}$ |
|---|---|---|---|
| 10-5 | 1 | 1 | 2 |
| 25-5 | 9 | 21 | 64 |
| 50-10 | 131 | 248 | 1,092 |
| 75-15 | 301 | 1,087 | 6,656 |

Table 2 shows the computation time for instances of model (21) in which each buffer receives exogenous arrivals. The coefficient of variation of the sample mean objective function for this version of the 75-15 test problem is about $10^{-4}$. We observe that for our larger systems, having an increased number of random parameters increases the computation time. Hence, in Table 2 we solve the test problems to a precision of up to $\epsilon = 10^{-3}$. These computations indicate that it is possible to solve the stochastic makespan problem for large networks with a desirable level of accuracy in reasonable time.

## 7. FUTURE WORK

There are two clear-cut ways in which the model considered in this article can be extended. In the current model, the percentage allocations are set before realizing the stochastic parameters, and cannot be changed afterward. One can envision a number of modifications to allow for a more flexible decision structure. One possible extension of this work is to consider the case in which a recourse action is possible; that is, the decision maker is allowed to change the allocations after some fixed time $T$. A more extensive modeling framework would allow recourse and stochastic changes in the parameters at a set of time points $T_1, \ldots, T_N$. Such a model is clearly a generalization of the model presented in this article, and Büke [9] explored this more general model. Another way to modify the decision structure is to allow the controller to select a *proportional allocation vector*, rather than a fixed allocation vector. When a proportional allocation is chosen, a station works full time on all fluid types with positive buffer levels.

It is also clear that for some applications, the makespan objective is not the most appropriate. The deterministic fluid model with a linear holding cost objective has been widely studied in the literature. One contribution of this article is to show that even the basic properties of the makespan problem change when the parameters are viewed as being random. Hence, the work herein raises the question of how the optimization characteristics of fluid problems under various objective functions change when random parameters are introduced.

## Acknowledgments

This research is supported in part by the National Science Foundation under grants DMI-0132038 and DMI-0217927.

## References

1. Anderson, E.J. & Nash, P. (1987). *Linear programming in infinite dimensional spaces*. New York: Wiley-Interscience.
2. Anick, D., Mitra, D. & Sondhi, M.M. (1982). Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal* 61(8): 1871–1894.
3. Atlason, J., Epelman, M. & Henderson, S. (2008). Optimizing call center staffing using simulation and analytic center cutting plane methods. *Management Science* 54: 295–309.
4. Avram, F., Bertsimas, D. & Ricard, M. (1995). Fluid models of sequencing problems in open queueing networks; an optimal control approach. In F.P. Kelly & R.J. Williams (eds.), *Stochastic networks*, IMA Volumes in Mathematics and its Applications, Vol. 71. New York: Springer-Verlag, pp. 199–237.
5. Bäuerle, N. (2001). Convex stochastic fluid programs with average cost. *Journal of Mathematical Analysis and Applications* 259(1): 137–156.
6. Bäuerle, N. (2001). Discounted stochastic fluid programs. *Mathematics of Operations Research* 26(2): 401–420.
7. Billings, R. (2003). A heuristic method for scheduling and dispatching of factory production using multiclass fluid networks. Ph.D. dissertation, University of Texas at Austin.
8. Birge, J. & Wets, R. (1986). Designing approximation schemes for stochastic optimization problems, in particular, for stochastic programs with recourse. *Mathematical Programming Study* 27: 54–102.
9. Büke, B. (2007). Optimal draining of fluid networks with parameter uncertainty. Ph.D. dissertation, University of Texas at Austin.
10. Chen, H. & Yao, D. (2001). *Fundamentals of queueing networks*. New York: Springer.
11. Dai, J.G. (1999). *Stability of fluid and stochastic processing networks*. MaPhySto Miscellanea Publication No. 9. Centre for Mathematical Physics and Stochastics.
12. Dai, J.G. & Weiss, G. (2002). A fluid heuristic for minimizing makespan in job shops. *Operations Research* 50(4): 692–707.
13. Dupačová, J. (1976). Minimax stochastic programs with nonconvex nonseparable penalty functions. In A. Prékopa (ed.), *Progress in operations research*. Eger, Hungary: Mathematica Societatis János Bolyai, pp. 303–316.
14. Edirisinghe, N. & You, G.-M. (1996). Second-order scenario approximation and refinement in optimization under uncertainty. *Annals of Operations Research* 64: 143–178.
15. Frauendorfer, K. (1988). Solving SLP recourse problems with arbitrary multivariate distributions: The dependent case. *Mathematics of Operations Research* 13: 377–394.
16. Frauendorfer, K. & Kall, P. (1988). A solution method for SLP recourse problems with arbitrary multivariate distributions: The independent case. *Problems of Control and Information Theory* 17: 177–205.
17. Gassmann, H. & Ziemba, W. (1986). A tight upper bound for the expectation of a convex function of a multivariate random variable. *Mathematical Programming Study* 27: 39–53.
18. Gürkan, G. (2000). Simulation optimization of buffer allocations in production lines with unreliable machines. *Annals of Operations Research* 93: 177–216.
19. Gürkan, G., Karaesmen, F. & Özdemir, Ö. (submitted). Optimal threshold levels in stochastic fluid models via simulation based optimization. *Discrete Event Dynamical Systems*.
20. Harrison, J.M. & Zeevi, A. (2005). A method for staffing large call centers using stochastic fluid models. *Manufacturing & Service Operations Management* 7: 20–36.
21. Huang, C., Ziemba, W. & Ben-Tal, A. (1977). Bounds on the expectation of a convex function of a random variable: With applications to stochastic programming. *Operations Research* 25: 315–325.
22. Iyengar, G. & Zeevi, A. (2008). Parameter uncertainty implications on asymptotic analysis and design of stochastic systems. Preprint.

23. Kall, P., Ruszczyński, A. & Frauendorfer, K. (1988). Approximation techniques in stochastic programming. In Y. Ermoliev & R.J.-B. Wets (eds.), *Numerical techniques for stochastic optimization*. Berlin: Springer-Verlag, pp. 33–64.
24. Kelley, J.E. (1960). The cutting plane method for solving convex programs. *SIAM Journal of Industrial and Applied Mathematics* 8: 703–712.
25. Kulkarni, V.G. (1997). Fluid models for single buffer systems. In J.H. Dshalalow (ed.), *Frontiers in queueing: Models and applications in science and engineering*. New York: CRC Press, pp. 321–338.
26. Madansky, A. (1959). Bounds on the expectation of a convex function of a multivariate random variable. *Annals of Mathematical Statistics* 30: 743–746.
27. Meyn, S.P. (2007). *Control techniques for complex networks*. Cambridge: Cambridge University Press.
28. Pullan, M.C. (1993). An algorithm for a class of continuous linear programs. *SIAM Journal of Control and Optimization* 31: 1558–1577.
29. Pullan, M.C. (1995). Forms of optimal solutions for separated continuous linear programs. *SIAM Journal of Control and Optimization* 33: 1952–1977.
30. Sethi, S.P., Yan, H., Zhang, H. & Zhang, Q. (2002). Optimal and hierarchical controls in dynamic stochastic manufacturing systems: A survey. *Manufacturing & Service Operations Management* 4(2): 133–170.
31. Shapiro, A. (2003). Monte Carlo sampling methods. In A. Ruszczyński & A. Shapiro (eds.), *Stochastic programming, handbooks in operations research and management science*. Amsterdam: Elsevier.
32. Sun, G., Cassandras, C. & Panayiotou, C. (2004). Perturbation analysis of multiclass stochastic fluid models. *Discrete Event Dynamical Systems* 14: 267–307.
33. van Slyke, R. & Wets, R. (1969). L-Shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal on Applied Mathematics* 17: 638–663.
34. Weiss, G. (1995). On optimal draining of fluid reentrant lines. In F.P. Kelly & R.J. Williams (eds.), *Stochastic networks*, IMA Volumes in Mathematics and its Applications, Vol. 71, New York: Springer-Verlag, pp. 93–105.
35. Weiss, G. (to appear). A simplex based algorithm to solve separated continuous linear programs. *Mathematical Programming A*.
36. Whitt, W. (2006). Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* 15(1): 88–102.